Decision model for coagulant dosage using genetic programming and multivariate statistical analysis for coagulation/flocculation at water treatment process

Sooyee Park*,†, Hyeon Bae, and Changwon Kim*****

*Waterworks Headquarter, Ulsan Metropolitan City, Ulsan 689-955, Korea
**Intelligent Control Systems Lab, Georgia Institute of Technology, USA
***School of Civil and Environmental Engineering, Pusan National University, Busan, Korea

(*Received 26 October 2007 • accepted 7 May 2008*)

Abstract−In this research, genetic programming and multivariate statistical analysis techniques have been applied for decision support on the coagulant dosage and the mixing ratio as two kinds of coagulants have been injected at the same time in the coagulating sedimentation process of water treatment. The coagulant dosage has typically been determined through the Jar-test, which requires a long experiment time in a field-water treatment plant. It is difficult to efficiently determine the coagulant dosage since water quality changes with time. As there are no human experts who have sufficient knowledge and experience in the field, coagulants may be injected with an improper mixing ratio, which causes poor performance in the coagulating sedimentation process. In this study, a model for the approximation of coagulant dosage has been developed using genetic programming (GP). The performance of this model was evaluated through validation. A guideline on the optimal mixing ratio between PACS (Poly Aluminum Chloride Silicate) and PAC (Poly Aluminum Chloride) has been provided through statistical analysis.

Key words: Coagulant Dosage, Genetic Programming, Multivariate Statistical Analysis, Water Treatment Plant

INTRODUCTION

The water treatment process has advanced to a tertiary water treatment process with the increase of water supply facilities and the combination with biological activated carbon adsorption (BAC). Together with this advance, automation technologies such as automatic monitoring of influent turbidity, conductivity, operating states of unit treatment process and effluent turbidity have improved significantly. Such improvements have made it possible to achieve efficient process management and stable effluent quality [1]. However, one area that the automation technology has not been applied to in the water treatment process is the determination of coagulant dosage. This determination is regarded as the most important one among various works that are conducted in the unit processes constituting the water treatment process [2,3], because it strongly affects the effluent quality. Nevertheless, the coagulant dosage is still determined every day with the operator's knowledge and manual methods. The type of coagulant and its dosage are affected by the properties of influent turbidity, alkalinity, conductivity, temperature, pH and so on. The Jar-test is performed according to the properties mentioned above, and the proper dosage is determined based on the result of this test.

The determination of the coagulant dosage using the Jar-test and the determination of injected-coagulant type based on the operator's knowledge are becoming a serious bottleneck for the automation of the water treatment process [4]. It was reported that such methods based on a human operator's knowledge have a disadvantage in that this knowledge couldn't be permanently guaranteed in the case of the retirement or relocation of the human expert [5]. The control of coagulant dosage with time is generally regarded as the urgent work since it must be conducted after the influent properties of the target water treatment plant is identified. In this paper, the model to determine the coagulant dosage was developed by using the genetic programming (GP) algorithm [6,7]. This model aims at performing decision support of the coagulant dosage at the water treatment plant. Because two kinds of coagulants - PAC and PACS - were injected together in the target water treatment plant, the most efficient mixing ratio was determined after analyzing the operating results of target water treatment plant, which were accumulated for two years. Finally, the information on the best mixing ratio was provided together with the predicted coagulant dosage in this paper. It is expected that the coagulant dosage will be determined with the aid of the decision-support model, which was developed in this research, without the need for the time-consuming Jar-test. It is also expected that rapid action on the influent variation will be taken using this developed decision support model. In addition, this developed model can be used as the tool for the proper and automatic coagulant injection in an emergency situation when there is no operator in the target water treatment plant.

MATERIAL AND METHOD

1. Data Collection

The data sets for 20 months (from 4-2004 to 12-2005) were collected from a full scale water treatment plant (WTP) for this research. The treatment capacity of the target WTP was 60,000 m^3/day, and the process of target WTP was tertiary treatment process with BAC. Fig. 1 shows the schematic diagram of chosen target WTP.

As shown in Fig. 1, it is most important to maintain the highest removal efficiency of the coagulating-sedimentation process for the efficient operation and good treatment performance of BAC, which

†To whom correspondence should be addressed.
E-mail: psy8515@hanmail.net

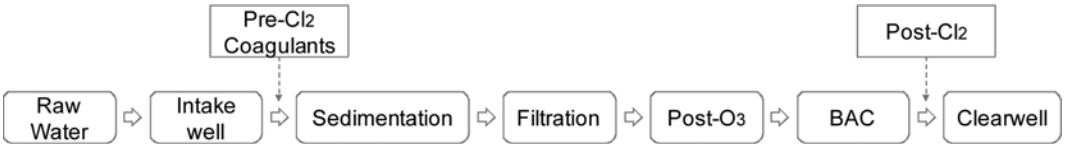


Fig. 1. Schematic flow diagram of the target water treatment system.

Table 1. Specification of the measuring instrument

Item	Model	Method	Range
pH	HDM-136A	Glass electrode	−1-14
Turbidity	TUF-100	Surface scattering	0-2,000 NTU
Alkalinity	ALF-100	Equivalence point titration	0-50/100 mg/L
Residual chlorine	CLF-110	Polarographic method	0-3 mg/L

Table 2. General characteristics of the used-coagulants

Item	PAC	PACS
pH	4.0	4.0
Specific gravity	1.32	1.36
Aluminum oxide (%)	10.06	16.35
Silicon dioxide (%)	-	0.30
Sulfate ion (%)	1.72	-
Fe (%)	0.0006	0.0028
As (mg/L)	0.014	0.114
Cr (mg/L)	0.28	0.20
Hg (mg/L)	0.009	0.024
Mn, Cd, Pb (mg/L)	0.00	0.00

is a post treatment process. As the variables that explain the properties of influent, the data on pH, conductivity and alkalinity, the dosage of PACS and PAC, were collected from the target WTP. And, the data on pH and turbidity were collected as the variables that explain the effluent properties of coagulating sedimentation process. Table 1 and Table 2 show the measuring methods and the typical properties of the PACS and PAC, which were used as coagulants, respectively. The time-variation of collected data is shown in Fig. 2. As shown in this figure, the period that only PACS was used was 300 days out of an operating periods of 629 days, which was about 48% of total operating period. And the period that only PAC was used was 270 days, which was about 39% of total operating period. The period that these two coagulants were injected together was 82 days, which was about 13% of total operating period. The influent properties were evaluated by the operator, and the coagulant type that should be used was determined based on the operator's knowledge and the amount of chemistry stock in the target water treatment plant. The coagulant dosage was determined with the Jar-test.

2. Genetic Programming (GP)

Genetic programming is a method that may evolve the model structure for calculating the target variable that should be approximated. Although it is an extension of the genetic algorithm, it is considered to be a more powerful tool. The fact that GP is based on global optimization makes the model prediction results better than other modeling approaches and seldom makes an over-fitted model [8]. The result of the GP is a single program, while that of genetic algorithm is a solution set that consists of numbers or symbols. The procedure for GP implementation is divided into four steps as follows [9,10].

(1) Generate an initial population of random compositions of the functions and terminals of the problem (computer programs).

(2) Execute each program in the population and assign it a fitness value according to how well it solves the problem.

(3) Create a new population of computer programs.

i) Copy the best existing programs

ii) Create new computer programs by mutation

iii) Create new computer programs by crossover (sexual reproduction)

(4) The best computer program that appears in any generation, the best-so-far solution, is designated as the result of genetic programming.

3. K-Means Clustering

Cluster analysis is a method to classify total objects into K number of groups or clusters with similar attributes, as there are N numbers of objects with several attributes. The clustering method is di-

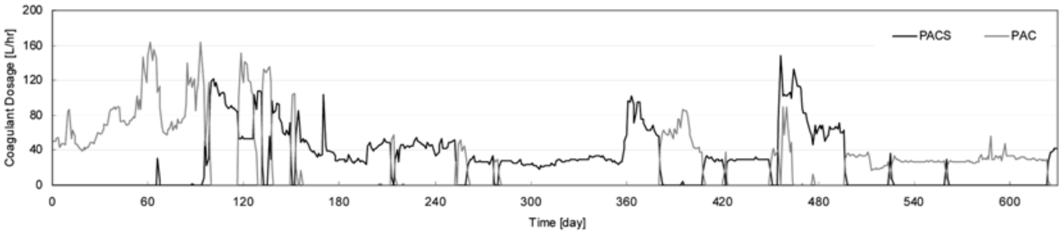


Fig. 2. Coagulant dosage at target water treatment plant.

vided into the hierarchical method and non-hierarchical method. The non-hierarchical method assigns each object to one of the K-independent clusters by defining the representative object or value of each cluster after the clusters are predefined as K-independent clusters. This assignment is not finished at once, but is repeated until the result of clustering analysis is up to convergence. K-means clustering analysis is a non-hierarchical method that is most generally used, and the K-means clustering procedure can be summarized as follows.

Step 0. (Selection of initial centroid) Select the coordinate of the K-object as the centroid of initial cluster with certain rules.

Step 1. (Assigning each object to each cluster) Assign each object to the nearest cluster after calculating the distances between each object and the centroids of the K-independent clusters.

Step 2. (Determination of a new centroid for each cluster) Determine the centroid of the new cluster.

Step 3. (Checking the convergence condition) After comparing the new centroid and the old one, if the result is within the convergence condition, the analysis is finished, if not, repeat step 1.

4. Factor Analysis

Some of common components, which can explain their correlations, can be derived even though so many variables in the multivariate data [11] are related complexly to each other, and the variable can be divided into the parts that reflect these common components and independent components. In factor analysis, the common component is called the common factor or simply the factor, while the independent component is called the unique factor. Factor analysis aims to derive the common factors from the correlation matrix for the variables, explaining the correlations between variables by using common factors, and to describe the property of each variable in a simply way. Therefore, the interpretation of the factor is conducted from the factor loading matrix or the factor structure matrix. As some particular variables in one factor have a high loading, it can be concluded that such variables are affected by one hidden factor together. Factor analysis is commonly used to derive the variables that are affected together by an unknown effect as the statistical correlation among variables is low.

RESULTS

1. Determination of Coagulant Dosage with Model Using GP

In order to develop a model that can approximate the total coagulant dosage, data from 629 days of operation were used. Among these data, 434 pieces of data for development and 195 pieces of data for validation were used by random selection. The GP, which was suggested by Koza (1992), was used for the development of the model and was coded by using Visual Studio C^{++} in this research. The variables that were used for the development of the model were influent turbidity, conductivity, alkalinity and temperature. The resulting equation consists of mathematical operations, which are predefined by a user. The equation is fitted based on the fitness function, so the equation has no physical meaning that can express phenomena of target systems. Fig. 3 shows the tree structure of the model in a genetic algorithm that was generated in this research. The parameters of genetic algorithm that were used in this paper are shown in Table 3. The probability, which was applied to the crossover and mutation, was determined by trial and error. The performance of the GP model was improved by using a random number from −2 to 2 as a constant.

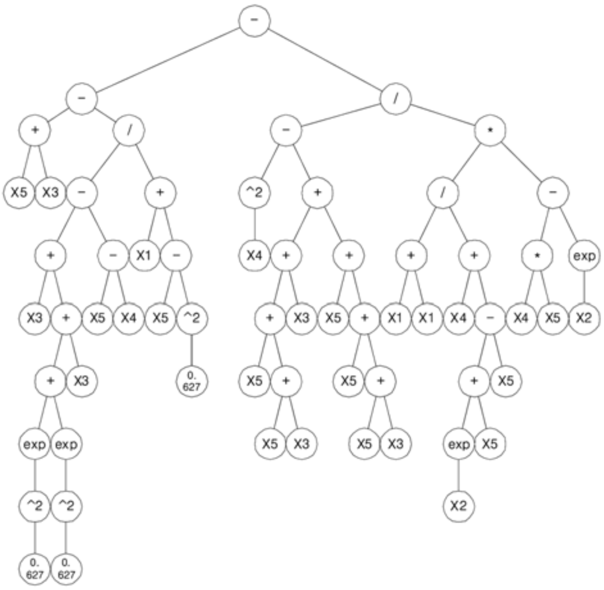

-
- /
+ / - *
X5 X3 - + ^2 + / -
+ - X1 - X4 + + + + * exp
X3 + X5 X4 X5 ^2 + X3 X5 + X1 X1 X4 - X4 X5 X2
+ X3 0.627 X5 + X5 + + X5
exp exp X5 X3 X5 X3 exp X5
^2 ^2 X2
0.627 0.627

Fig. 3. Generated model as tree structure for estimating coagulant dosage by genetic programming using influent variables - x_1 (pH), x_2 (conductivity), x_3 (alkalinity), x_4 (temperature), x_5 (turbidity).

Table 3. Parameters used at genetic programming

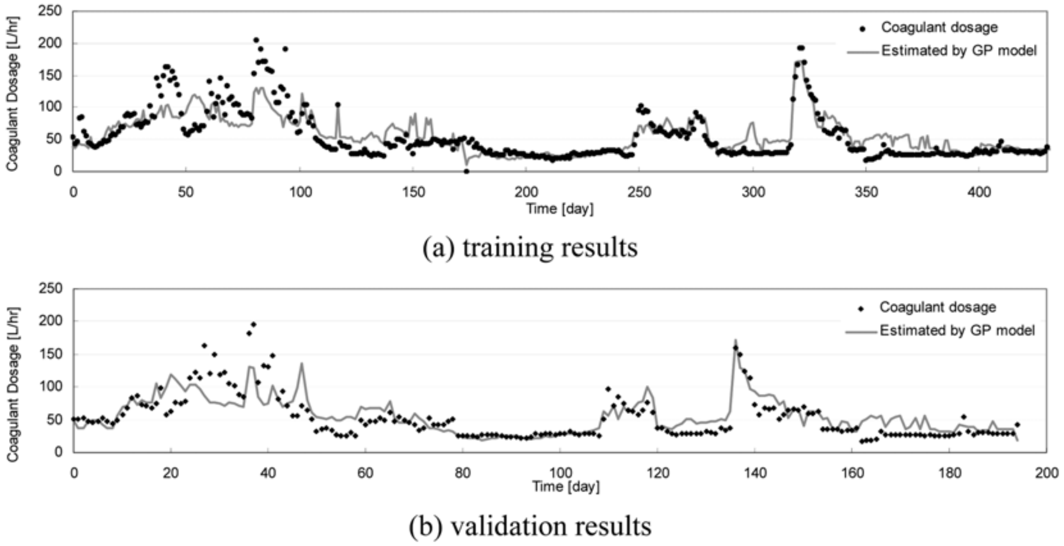
Parameters	Networks setting	Parameters	Networks setting
Function set	+, –, *, /, pow, exp	Individual number	1,000
Constant value	–2.000-2.000	Probability of crossover	0.9
Initial depth	8	Probability of mutation	0.008
Create depth	5		

The performance of the GP model after training and validation is shown in Fig. 4. The model performance was evaluated with RMSE (Eq. (1)). It was confirmed that the RMSE for training and validation were 24.21 and 20.84, respectively. In general parameter modeling, RMSE of training is lower than that of testing, because models can be trained to reduce errors corresponding to applied data. In other words, the model should be the best for the training data, but GP is a different type of modeling methods. GP is not parameter-based modeling, but structure-based modeling. Models generated by GP are tuned to the principal trend of the applied data. That is, GP can generate global models that can avoid an over-fitting problem. Therefore, RMSE of testing is often smaller than that of training.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (1)$$

2. Determination of Mixed-injection Ratio for Coagulants with Statistical Analysis

There are two variables aside from total coagulant dosage that must be considered for the operation of the coagulating sedimentation process in target water treatment plant. One is the determination of the coagulation type if only one coagulant is used. Another



(a) training results

(b) validation results

Fig. 4. GP model training and validation results.

is the determination of the mixing ratio if two coagulants are used together. Although an expert determines the coagulant type and mixing ratio based on the amount of coagulant stock and influent properties, if there are certain guidelines for doing this procedure, the guidelines will play an important role when a non-expert operates the target water treatment plant or if automation of the coagulant injection system is necessary.

The data that were collected for 20 months were classified into five groups by K-means clustering that focused on the records of coagulant injection. The centroid of each classified group is shown in Table 4. It was observed that the classifying injection ranges into five groups and were appropriate when the injection ranges were classified based on the centroid of each group. However, the mean of the clusters calculated by k-means was not clearly expressed to be used as the significant index, which is applied to handle the control variables, so in this study, the range was modified to 0, 1, 0-0.3, 0.3-0.7, and 0.7-1. The data sets that consisted of the influent turbidity and effluent turbidity could be classified according to the five efficiencies, and the removal efficiencies could be calculated based on the classified groups. Table 5 shows removal efficiencies of each cluster.

Table 4. K-mean clustering results and modified numerical range of cluster

Item	Cluster no. (Case no.)				
	1	2	3	4	5
Mean of cluster	0.001	0.269	0.603	0.790	1.000
Modified range of cluster	0.0	0.0-0.3	0.3-0.7	0.7-1.0	1.0

Table 5. Removal efficiencies of each cluster

Case no.	Description	Influent turbidity (NTU)	Effluent turbidity (NTU)	Average of efficiency (%)
1	PACS only	18.02	0.60	94.38
2	PAC only	14.52	0.63	93.73
3	PACS and PAC	44.57	1.03	95.79
3-1	PACS/Total<=0.3	25.85	0.76	95.23
3-2	0.3<PACS/Total<0.7	69.88	1.24	95.98
3-3	0.7<PACS/Total<=1	46.19	1.18	96.26

As shown in these tables, it was confirmed that the mixed-injection of the two kinds of coagulants was conducted for high influent turbidity, and the best performance was shown as the injection ratio of the PACS to total coagulant dosage was above 0.7. The mixed-injection ratio of PACS and PAC was 0.3-0.7 for influent with high turbidity, about 60 NTU. The removal efficiency was 95.98%, which was better than when only one coagulant was injected. As the influent turbidity approached 45 NTU, the injection ratio of PACS was above 0.7. At this time, the removal efficiency was 96.26%, which was the highest efficiency.

Since it was difficult to identify the cause-effect relationship of the coagulant dosage and injection ratio by only considering simple mean values, factor analysis for the exploration of the potential structure, which was contained in the multivariate data, was conducted. Through this, the correlation of each factor was investigated. As

Table 6. Factor matrix for each variables by factor analysis

Item	Component			
	1	2	3	4
pH of influent	−0.160	0.787	−0.331	−0.087
Conductivity of influent	0.108	0.801	0.143	0.038
Alkalinity of influent	0.145	0.705	0.053	0.112
Temperature of influent	0.261	−0.360	0.680	0.096
Turbidity of influent	0.843	−0.053	0.214	−0.029
Total coagulant dosage	0.628	0.231	0.647	−0.036
Turbidity of effluent	0.773	−0.107	0.243	−0.060
pH of effluent	−0.399	0.740	0.084	−0.103
Chlorophyll - a of effluent	−0.037	0.026	−0.011	0.981
PACS dosage	0.841	0.040	−0.340	0.039
PAC dosage	−0.083	0.198	0.939	−0.069

result of factor analysis, four factors with an eigenvalue that explains more than one variable were extracted. Table 6 shows the factor matrix, which was extracted by using factor analysis. This table shows that the first factor is affected by influent turbidity, total coagulant dosage, effluent turbidity, and PACS dosage. This can tell us that the most important factor to determine the coagulant dosage is the influent turbidity, and that influent turbidity and coagulation dosage affect the effluent turbidity of coagulating sedimentation process. It can also tell us that the amount of PACS in the total coagulant dosage is an important variable to determine the effluent turbidity. Therefore, it could be discerned that its result corresponded with the interpretation on the performance based on the mean value of efficiency, which was conducted previously. The second factor is related to influent properties and effluent pH. This shows that influent pH, conductivity and effluent pH can be explained as the same factor. This relation is also accepted as an appropriate one in the aspects of water quality, and this result tell us that the data that were used for this research reflect the properties of target influent and field processes. Finally, it is confirmed that the third factor is affected by the variables related to temperature and total coagulant dosage, and the fourth factor is affected by only chlorophyll concentration.

CONCLUSION

A model to predict the coagulant dosage and the guideline on the injection ratio of various coagulants were developed with the purpose of providing decision support for the determination of coagulant dosage and injection ratio at the coagulating sedimentation process, which is regarded as the main treatment process in a water treatment plant. A model that can approximate the total coagulant dosage, as the variables related to influent qualities were measured, was developed by using GP. It was observed that the error of the developed model was about 20.84. Furthermore, the removal efficiency of each group after classifying the coagulant injection ratio into five groups with K-means clustering was estimated so that non-experts could determine the proper injection ratio of the coagulants easily, since the two kinds of coagulant were injected at the same time through mixing at the target water treatment plant. As a result, it was confirmed that the highest removal efficiency was shown to be above 0.7 for the mixed-injection ratio of PACS, and the mixing ratio of PACS and PAC was 0.3-0.7 for the high influent turbidity. The results of factor analysis, which was conducted to interpret the factors that were important in the collected data, told us that the PACS dosage was affected by influent turbidity, and its dosage affected the effluent turbidity significantly. It could be discerned that what should be checked for high effluent turbidity was the amount of PACS that could be used, and its injection ratio must be above 0.3 regardless of the problems related to the price and the amount of stock. The decision support and automation of a water treatment plant is necessary and that is related to the price of drinking water. If useful information can be extracted from the accumulated data and provided to non-experts who do not have much knowledge and experience, or if the coagulant can be injected automatically in an emergency situation of a situation when there are no operators, it would be very helpful for the efficient usage of operators and economical operation of the treatment process.

REFERENCES

1. S. H. Kim and J. Y. Yoon, *J. Wat. Suppl.: Res. & Technol.-AQUA*, **54**, 95 (2005).
2. L. Lu, H. Ratnaweera and O. Lindholm, *VATTEN*, **59**, 227 (2003).
3. Y. Zhan, *China Water and Wastewater*, **20**, 55 (2004).
4. H. Bae, S. Kim and Y. Kim, *Lecture Notes in Computer Science, Springer-Verlag GmbH*, **3735**, 371 (2005).
5. Y. Kim, Ph. D. Thesis, Pusan National University, Busan (2006).
6. X. Cai, D. C. McKinney and L. S. Lasdon, *Advances in Water Resources*, **24**, 667 (2001).
7. J. H. Cho, K. S. Sung and S. R. Ha, *J. of Environmental Management*, **73**, 229 (2004).
8. L. M. Deschaine, J. McCormack, D. Pyle and F. Francone, *PC AI*, **15**, 35 (2001).
9. J. R. Koza, in *Genetic programming : On the programming of computers by means of natural selection*, MA: The MIT Press, Cambridge (1992).
10. M. Walker, in *Introduction to genetic programming*, Technical report (2001). http://www.rmltech.com/
11. J.-B. Serodes, M. J. Rodriguez and A. Ponton, *Environmental Modelling and Software*, **16**, 53 (2001).